

PB87-114591

Compartment Model Approaches for Estimating the
Parameters of a Chronic Disease Process under
Changing Risk Factor Exposures

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1986



TECHNICAL REPORT DATA

(Please read instructions on the reverse before complete)

1. REPORT NO. EPA/600/J-86/162		3. PB87-114591	
4. TITLE AND SUBTITLE COMPARTMENT MODEL APPROACHES FOR ESTIMATING THE PARAMETERS OF A CHRONIC DISEASE PROCESS UNDER CHANGING RISK FACTOR EXPOSURES (Journal Version)		5. REPORT DATE 1986	
7. AUTHOR(S) Kenneth G. Manton*, Eric Stallard*, John P. Creason, Wilson B. Riggan, and Max A. Woodbury+		6. PERFORMING ORGANIZATION CODE	
9. PERFORMING ORGANIZATION NAME AND ADDRESS EPA,HERL,ITD,TB, MD-82, RTP, NC 27711 *Center for Demographic Studies, Duke University Durham, NC 27706 +Department of Community and Family Medicine, Duke U.		8. PERFORMING ORGANIZATION REPORT NO.	
12. SPONSORING AGENCY NAME AND ADDRESS HEALTH EFFECTS RESEARCH LABORATORY - RTP, NC OFFICE OF RESEARCH AND DEVELOPMENT U.S. ENVIRONMENTAL PROTECTION AGENCY RESEARCH TRIANGLE PARK, NC 27711		10. PROGRAM ELEMENT NO.	
		11. CONTRACT/GRANT NO.	
		13. TYPE OF REPORT AND PERIOD COVERED	
		14. SPONSORING AGENCY CODE EPA/600/11	
15. SUPPLEMENTARY NOTES Computers and Biomedical Research 19, 151-169 (1986) Feb.			
16. ABSTRACT Compartment model approaches have been proposed for the analysis of the age incidence of specific types of cancer. These models represented the age increase in incidence as the result of a compound hazard function where individual level risks were described by the Weibull hazard function and where the population level hazard rate is a continuous mixture of the Weibull hazards. These formulations assumed that the mixing function, which described differences in risk due to different exposure histories, was constant after the age at which the model was first applied. In this paper we show how the mixing distribution can be allowed to change with time reflecting changing exposures. The model is fitted to U.S. lung cancer mortality data where for recent male cohorts there appears to be changing patterns of exposure possibly related to recent declines in male smoking. The implications for future lung cancer mortality trends in the U.S. are discussed.			
17. KEY WORDS AND DOCUMENT ANALYSIS			
a. DESCRIPTORS		b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC		19. SECURITY CLASS (This Report) UNCLASSIFIED	21. NO. OF PAGES 20
		20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

Compartment Model Approaches for Estimating the Parameters of a Chronic Disease Process under Changing Risk Factor Exposures

KENNETH G. MANTON,* ERIC STALLARD,* JOHN P. CREASON,†
WILSON B. RIGGAN,† AND MAX A. WOODBURY‡

*Center for Demographic Studies, Duke University, Durham, North Carolina 27706;

†U.S. Environmental Protection Agency, Research Triangle Park,
Durham, North Carolina 27711; and ‡Department of Community and Family Medicine,
Duke University Medical Center, Durham, North Carolina 27710

Received May 14, 1985

Compartment model approaches have been proposed for the analysis of the age incidence of specific types of cancer. These models represented the age increases in incidence as the result of a compound hazard function where individual level risks were described by the Weibull hazard function and where the population level hazard rate is a continuous mixture of the Weibull hazards. These formulations assumed that the mixing function, which described differences in risk due to different exposure histories, was constant after the age at which the model was first applied. In this paper we show how the mixing distribution can be allowed to change with time reflecting changing exposures. The model is fitted to U.S. lung cancer mortality data where for recent male cohorts there appear to be changing patterns of exposure possibly related to recent declines in male smoking. The implications for future lung cancer mortality trends in the United States are discussed. © 1986 Academic Press, Inc.

I. INTRODUCTION

Statistical analysis of chronic disease processes must reflect the fact that such processes evolve over a lengthy period of time and through various disease stages. One approach suggested for the analysis of such processes is that of stochastic compartment models (1-3). In constructing these models, one (a) specifies a set of discrete health states, (b) specifies specific parametric hazard functions to describe transitions between those states, and (c) fits the marginal hazard function to age- and time-specific morbidity or mortality data.

One specific stochastic compartment model of cancer utilizes the Weibull hazard function to describe the age dependence of cancer incidence at the individual level and a gamma mixing distribution for the proportionality constant in the Weibull function to represent individual differences in risks. Individual differences in risk may arise as a function of either intrinsic factors or

exposure to risk factors. For that component of individual risk differences due to risk factor exposure it is important to show how risk covariates can affect the gamma mixing distribution and to show how parameter estimation can be conducted, given that the gamma mixing distribution changes through time to represent changes in individual risks due to changing exposures.

This paper describes how the gamma mixing distribution relates to observable risk factors and how observable risk factors can be brought into the model. Also shown is how the gamma mixing distribution for selected cohorts can be allowed to change over time. These procedures are applied to U.S. lung cancer mortality data for the period 1950 to 1977 where recent male cohorts have exhibited rapidly changing risk distributions, probably due to changing smoking patterns.

II. THE MODEL

This section describes four features of a general compartment model of cancer risks which can describe changing susceptibility and covariate effects. First, the basic probability structure of the compartment model is presented. This development does not assume the prior selection of a specific hazard or mixing function. Second, a specific compartment model is developed for lung cancer based upon clinical and epidemiological evidence. Third, it is shown how the model can be estimated for the case that risk factor exposure changes over time. Finally, the relationship of the gamma mixing distribution and measured covariates is derived.

A. Transitions in a Heterogeneous Population

To develop our compartment model, we first consider a generalization of the standard homogeneous population survival model to the case where death, failure, or any other specified transition is dependent upon a variable summarizing individual differences in risk, say Z . Let $\{(X, Z)\}$ denote a set of bivariate random vectors which are independently and identically distributed, having the cumulative distribution function $Q(z)$ on Z and, conditionally on $Z = z$, $F(x|z)$ on X . The variate X represents the random time at which a transition occurs; in our analysis of mortality, $X = x_i$ is the observed lifetime of the i th individual, which is assumed to depend on an unobserved value of the covariate Z . We assume that $F(x|z)$ is absolutely continuous on $(0, \infty)$, with $F(0|z) = 0$ and $F(\infty|z) = 1$ and that $Q(z)$ is absolutely continuous on $(-\infty, \infty)$ with $Q(-\infty) = 0$ and $Q(\infty) = 1$. The conditional survival function is defined as

$$S(x|z) = 1 - F(x|z) \quad x \geq 0. \quad [1]$$

The conditional survival function, $S(x|z)$, the conditional hazard rate, $h(x|z)$, and the conditional probability density $f(x|z)$, are interrelated as

$$S(x|z) = \exp \left\{ - \int_0^x h(t|z) dt \right\} \quad [2]$$

$$h(x|z) = \frac{-1}{S(x|z)} \frac{\partial}{\partial x} S(x|z) \quad [3a]$$

$$= f(x|z)/S(x|z) \quad [3b]$$

and

$$f(x|z) = \frac{-\partial}{\partial x} S(x|z) \quad [4a]$$

$$= h(x|z)S(x|z). \quad [4b]$$

The marginal functions $S(x)$, $f(x)$, and $h(x)$ are the expected values of the corresponding conditional functions $S(x|z)$, $f(x|z)$, and $h(x|z)$. For the first two of these, $S(x)$ and $f(x)$, the expectation is taken over the distribution of Z at $x = 0$. For $h(x)$, however, the expectation must be conditional on survival to age x . From elementary principles, it follows that

$$S(x) = \int_{-\infty}^{\infty} S(x|z)q(z)dz, \quad [5]$$

where $q(z)$ is the probability density of Z at initial age $x = 0$, i.e.,

$$q(z) = \frac{d}{dz} Q(z). \quad [6]$$

Equation [5] expresses the marginal survival function $S(x)$ as the average of the survival functions $S(x|z)$ in a heterogeneous cohort. By differentiating Eq. [5] with respect to x , changing sign, and replacing the derivatives of $S(x)$ and $S(x|z)$ (see Eq. [4a]) we obtain

$$f(x) = \int_{-\infty}^{\infty} f(x|z)q(z)dz, \quad [7]$$

which expresses the marginal probability density function $f(x)$ as the average of the conditional density functions $f(x|z)$ in a heterogeneous cohort.

If we replace $f(x)$ and $f(x|z)$ in Eq. [7] (see Eq. [4b]) and divide both sides of the resulting expression by $S(x)$, we obtain

$$h(x) = \int_{-\infty}^{\infty} h(x|z)[S(x|z)q(z)/S(x)]dz, \quad [8a]$$

which can be rewritten as

$$h(x) = \int_{-\infty}^{\infty} h(x|z)q(z|X > x)dz, \quad [8b]$$

where $q(z|X > x)$ is the conditional density of Z given survival to age x , where X is the random time to death, failure, or any other specified transition. Thus Eq. [8b] establishes the following result: The marginal hazard function $h(x)$ is the expected value of the individual forces of mortality $h(x|z)$ among the cohort members who survive to age x or older.

Equation [8b] requires that we know the conditional density $q(z|X > x)$,

given survival to age x or older. An alternative expression for $h(x)$ which requires only the marginal density $q(z) = q(z|X > 0)$ follows from Eq. [5]. Specifically, replacing $S(x)$ and $S(x|z)$ we obtain

$$\exp \left\{ - \int_0^x h(t) dt \right\} = E_Z \left[\exp \left\{ - \int_0^x h(t|z) dt \right\} \right]. \quad [9]$$

By logarithmically transforming [9] and differentiating with respect to x we obtain:

$$h(x) = \frac{-\partial}{\partial x} \ln \left\{ E_Z \left[\exp \left\{ - \int_0^x h(t|z) dt \right\} \right] \right\}. \quad [10]$$

It may happen that Eq. [10] is easier to evaluate than [8b].

B. A Model of Cancer Mortality

In this section we will consider specialization of the heterogeneous population survival model to the proportional hazard form (4):

$$h(x|z) = z\lambda(x) \quad z \geq 0, \quad [11]$$

where $\lambda(x)$ is a standard force of mortality at age x for $z = 1$, and where z is strictly nonnegative. Our development proceeds as if only one cause of death were operating on the population. This is done to simplify the exposition. It follows from Eq. [10] that the same results obtain in the presence of J causes of failure, each with its own cause specific frailty or susceptibility variate Z_j , $j = 1, \dots, J$, if (a) the J frailty variates are statistically independent and (b) $h_j(x|z_j)$ is defined as the hazard rate at age x for the select cause given that one survives the other $J - 1$ causes to at least age x (5).

The proportional hazard model in Eq. [11] can be rewritten with z represented as a product of $K + 1$ effects in the form,

$$z = g_0(y_0)g_1(y_1) \dots g_K(y_K), \quad [12]$$

where $y_0, y_1, \dots, y_K$ are independent risk factors and $g_0, g_1, \dots, g_K$ are a set of nonnegative dose-response functions. Substitution of Eq. [12] for z in [11] yields an expression which is linear under logarithmic transformation and which can be generalized to include risk factor interactions. The inclusion of age by risk factor interactions must be carefully evaluated, however, since age is a time measure and mortality selection on the population operates over the time dimension. One strategy which can be used to include age by risk factor interactions in an analysis is to estimate separate risk factor coefficients for each age strata (e.g., as in Ref. (6)).

In this section we regard Eq. [12] as implicit in order to focus on biological assumptions leading to specification of $\lambda(x)$ and $q(z|X > x)$ under the assumption that z is unobserved. We will make assumptions about the parametric functional forms of two types of functions: (a) the standard force of mortality function $\lambda(x)$ in [11]; and (b) the marginal density function $q(z)$ of Z . With assumptions about both types of functions there are sufficient constraints to

perform the integration in [8a] to obtain the marginal force of mortality $h(x)$, which is the expected death rate in the age interval $(x - \frac{1}{2}dx, x + \frac{1}{2}dx)$.

In specifying a model of the standard force of mortality for deaths due to cancer, we need to combine aspects of biological insight into three phases of the natural history of the cancer mortality processes: (a) cellular transformation, or the initiation of tumor growth; (b) the growth phase itself; and (c) the clinical, treatment phase.

Although the precise biochemical changes in the DNA of the precursor to a malignant cell have yet to be described, the statistical laws describing the rate of passage of a population of cells through successive stages of this process are generally accepted as following the Weibull hazard form, increasing as some integral power of age (7-18). Hence, the rate of incidence of diagnosed cancer will be described by the age curve:

$$\lambda(x) = \alpha(x - l)^{m-1}, \quad [13]$$

In Eq. [13] m represents the requisite number of transformations or mutations in the cell nucleus for malignant growth to initiate, $(x - l)$ represents the time over which the transformations occur, l represents the tumor growth time, and α is a parameter which is proportional to the product of the m transformation probabilities. Ref. (10) discusses the effect of constant exposure to carcinogens under the multistage model and shows that if any one of several stages is dose dependent the relative risk in the exposed group will be constant and independent of age. Thus, a multiplicative interaction of the form in Eq. [12] is predicted if the $K + 1$ risk factors affect $K + 1$ distinct stages in the carcinogenic process.

The parametric function in Eq. [13] has also been found to describe the age curve of cancer mortality for a broad range of cancer sites (e.g., Ref. (13)). Thus one can also use Eq. [13] as a model of cancer mortality where (a) l is now the sum of the tumor growth time and the time between diagnosis and death, and (b) α is now reduced by the fraction of cancer incidence which is successfully treated. If both components are relatively constant over time, then the relative trends in α derived from mortality data will reflect the relative trends in cancer incidence (3). In the following, we assume that this is the case.

Substituting Eq. [13] for $\lambda(x)$ in [11] gives, as the model of individual cancer mortality:

$$h(x|z) = z \alpha (x - l)^{m-1}, \quad [14]$$

from which it is obvious that z is a measure of relative risk, but with respect to the standard force of mortality, *not* to the population force of mortality (i.e., not to $h(x)$). If the z values are allowed to change over time due to risk factor exposures, then the changes in the mean can be absorbed into α to maintain z as a measure of relative risk (see next section).

If we use $\Lambda(x)$ to denote the integral of $\lambda(x)$ over the interval 0 to x , then Eq. [11] implies

$$\int_0^x h(t|z) dt = z \Lambda(x) \quad [15]$$

in which case the right-hand sides of Eqs. [9] and [10] simplify to:

$$S(x) = \mathcal{L}[\Lambda(x)] \quad [16]$$

and

$$h(x) = \frac{-\partial}{\partial x} \ln\{\mathcal{L}[\Lambda(x)]\}, \quad [17]$$

where $\mathcal{L}[\cdot]$ denotes the Laplace transform of $q(z)$. Thus, Eq. [17] shows that, for proportional hazard models of the form [11], $q(z)$ must be restricted to the set of probability density functions for which z is nonnegative and for which the Laplace transform exists.

For reasons to be discussed below, we assume $q(z)$ is the gamma density, so that Eq. [16] becomes:

$$S(x) = \{\beta/[\beta + \Lambda(x)]\}^\gamma \quad \beta > 0, \gamma > 0, \quad [18]$$

where β^{-1} and γ are the scale and shape parameters of the gamma distribution; and Eq. [17] becomes

$$h(x) = \gamma\lambda(x)/[\gamma + \Lambda(x)], \quad [19]$$

where we have eliminated β from Eq. [19] by rescaling z to have an expectation of 1. For the gamma density, this means that the shape and scale are reciprocal since

$$\mathcal{E}_Z\{z\} = \gamma/\beta \text{ and } \text{var}_Z(z) = \gamma/\beta^2. \quad [20]$$

By replacing $\lambda(x)$ in (19) with the Weibull function in Eq. [14] one obtains the hazard function for the population (1)

$$h(x) = \alpha(x-l)^{m-1}/[1 + \alpha(x-l)^m/m\gamma], \quad [21]$$

with parameters α , m , γ , and l . Let n_i denote the number of cancer deaths in a given year at nominal age x_i (= age at last birthday) and let N_i denote the midyear population at nominal age x_i ; then the likelihood is approximated by

$$L = \prod_i [N_i h(x_i + \frac{1}{2})]^{n_i} \exp[-N_i h(x_i + \frac{1}{2})], \quad [22]$$

where $h(x_i + \frac{1}{2})$ is parameterized with the right side of Eq. [21] (or [25], [34]; see below), with $x \leftarrow x_i + \frac{1}{2}$. Parameter estimation can be conducted using maximum likelihood procedures of the type reported in Ref. (1).

C. Changing Risk Factor Exposure

The previous section showed how a model of cancer incidence or mortality could be developed from the gamma mixture of individual level Weibull processes. The gamma mixing distribution was selected for several reasons. First, the gamma distribution has been found acceptable by many researchers in similar types of problems (19-23). Second, the gamma distribution is flexible

and capable of adjusting to a wide range of shapes from an extreme right skewed L-shape through the familiar exponential form, up to the approximating normal form. Third, in many epidemiological studies it has been found that the distribution of risks to specific chronic diseases will be right skewed with a long tail. This seems particularly true for diseases whose risks are affected by exposures to toxic substances where the bulk of the population will have low exposures and a small proportion of the population will have high, but continuously varying exposures (e.g., smoking, asbestos). Such distributions could also result in diseases with multiple genetic determinants of risk where the probability of each genetic factor occurring is independent. Furthermore, as we will demonstrate later in this paper, the gamma distribution can be seen to be consistent with the accumulation of the effects of many covariates, each of which follows a general two-parameter beta distribution.

An important computational property of the gamma distribution is that it is closed under mortality selection so that it is possible to deal with any number of arbitrary age intervals knowing that the distribution of susceptibility at the start of each arbitrary interval will be the gamma distribution. By way of contrast, the rectangular distribution also possesses a Laplace transform but is not closed under mortality selection. Hence, if the distribution of susceptibility is rectangular at birth, it will not be rectangular at any subsequent age.

Ref. (24) showed that the closure of the gamma model under mortality selection applied to the entire class of natural exponential family distributions and suggested the inverse Gaussian distribution as a plausible alternative to the gamma. Ref. (25) evaluated the different effects of the gamma versus the inverse Gaussian model on the estimated parameters of four types of hazard functions (including the Weibull function) and found that the gamma models fit both total mortality and lung cancer mortality data better than the corresponding inverse Gaussian models.

The conditional gamma density, given survival to age x , can be written in the following form

$$q(z|X > x) = [z\beta_x]^\gamma \frac{\exp[-z\beta_x]}{z\Gamma(\gamma)} \quad [23]$$

where

$$\beta_x = \beta + \Lambda(x). \quad [24]$$

Clearly $\beta_0 = \beta$. The increases in β_x over age reflect the effects of mortality selection. Thus the shape parameter γ is available to model the effects of changes in risk factor exposures. In the next section we will show how covariates which are beta distributed could compose to yield a sequence of gamma distributions with different γ values. Here, we will consider a more general model where γ in Eq. [23] is replaced by γ_x , and α is replaced by α_x , subject to the conditions that γ_x is proportional to α_x and that $\gamma_0 = \gamma$ and $\alpha_0 = \alpha$. In this case, it follows from Eqs. [14] and [23] that the marginal hazard, $h(x)$, becomes

$$h(x) = \alpha_x(x-l)^{m-1}/[1 + \rho(x-l)^{m/m}], \quad [25]$$

where the constant of proportionality, ρ^{-1} , is defined by the relationship that, for all x ,

$$\gamma_x = \alpha_x / \rho \quad [26a]$$

$$\beta_x = \beta_0 [1 + \rho(x - l)^m / m]. \quad [26b]$$

Note that Eq. [25] is a proportional hazard model if all covariate effects are assumed to operate through α_x . In our application we will model α_x as a step function which changes at integral age boundaries. It is also possible to model α_x as a parametric function of age.

D. The Effects of Covariates on $Q(z)$

The assumption that the heterogeneity variate Z is gamma distributed permits certain qualitative statements to be made about the effects of unobserved covariates. This section develops an idea suggested by Ref. (24) that if the dose-response functions g_k in Eq. [12] are proportional to an appropriate family of beta variates, then the effect of a given g_k is to transform the distribution of Z from one gamma form to another. Specifically, assume that: (a) the dose-response functions in Eq. [12], $g_1, \dots, g_K$ are potentially observable but that g_0 is inherently unobservable; (b) each function is scaled so that its expected value is unity; and (c) all functions are statistically independent. Let G_k denote the k th variate function and let the notation $G_0 \sim \Gamma(1, \gamma_0)$ indicate that G_0 is gamma distributed with a mean of 1 and a shape parameter γ_0 . Hence, the density function for G_0 can be written as

$$q(g_0) = (g_0 \gamma_0)^{\gamma_0} \frac{\exp(-g_0 \gamma_0)}{g_0 \Gamma(\gamma_0)}, \quad (\gamma_0 > 0) \quad [27]$$

where $\Gamma(\gamma)$ denotes the gamma function defined by

$$\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} \cdot \exp(-x) dx. \quad [28]$$

Let the notation $G_k \sim B(1, \delta_k, \theta_k)$, $k = 1, \dots, K$, indicate that G_k is beta distributed with a mean of 1, and shape parameters δ_k and θ_k . Hence, the density function for G_k can be written as

$$q(g_k) = \left(\frac{\delta_k}{\delta_k + \theta_k} \right)^{\delta_k + \theta_k - 1} \frac{g_k^{\delta_k - 1} [(\delta_k + \theta_k)/\delta_k - g_k]^{\theta_k - 1}}{B(\delta_k, \theta_k)} \quad (\delta_k > 0, \theta_k > 0) \quad [29]$$

where g_k is restricted to the range 0 to $(\delta_k + \theta_k)/\delta_k$, and where $B(\delta, \theta)$ denotes the beta function defined by

$$B(\delta, \theta) = \int_0^1 x^{\delta-1} (1-x)^{\theta-1} dx \quad [30]$$

$$= \Gamma(\delta) \cdot \Gamma(\theta) / \Gamma(\delta + \theta). \quad [31]$$

Let Z_k denote the heterogeneity variate with the first $k + 1$ functions unobserved, i.e.,

$$z_k = \prod_{i=0}^k g_i, \quad [32]$$

so that $z_0 = g_0$ and $z_1 = g_0 \cdot g_1$. A straightforward calculation shows that if $G_1 \sim B(1, \gamma_1, \gamma_0 - \gamma_1)$ and $Z_0 \sim \Gamma(1, \gamma_0)$, then $Z_1 \sim \Gamma(1, \gamma_1)$. Since the shape parameters γ_0 and γ_1 are the inverses of the squared coefficient of variation, c_0^2 and c_1^2 , the expected increase in this statistic due to the effect of G_1 is the difference: $c_1^2 - c_0^2 = \gamma_1^{-1} - \gamma_0^{-1}$. The squared coefficient of variation for the beta variate $G \sim B(1, \delta, \theta)$ is obtained from

$$c^2 = \frac{\theta}{\delta(\delta + \theta + 1)}. \quad [33]$$

If each $G_k \sim B(1, \gamma_k, \gamma_{k-1} - \gamma_k)$, then each $Z_k \sim \Gamma(1, \gamma_k)$. Thus, the assumption that Z_K is gamma distributed in the model without covariates can be rationalized in terms of a model in which the conditional distribution of Z_K given covariates (i.e., the distribution of $Z_k, k < K$) is also of gamma form but with a larger shape parameter or, equivalently, a smaller coefficient of variation.

It can also be shown that the products of the G_k s obey the property $G_k \cdot G_{k+1} \sim B(1, \gamma_{k+1}, \gamma_{k-1} - \gamma_{k+1})$. By repeated application of this property we find that $G_1 \cdots G_K \sim B(1, \gamma_K, \gamma_0 - \gamma_K)$, so that $Z_0 \sim \Gamma(1, \gamma_0)$ implies that $Z_K \sim \Gamma(1, \gamma_K)$. Thus, the multiplicative effects of several independent covariate functions can be combined into one beta variate which operates on Z_0 to produce Z_K , both of which are gamma distributed variates.

III. APPLICATIONS

The model presented above was applied to data on respiratory cancer mortality among white males in the United States over the period 1950 to 1977. This section is in two parts. First, we describe the characteristics of these data and summarize findings of certain previous analyses and modeling efforts based on these data. This discussion motivates the current effort to allow the mixing distribution to change with time and to investigate the effects of unobserved covariates by stratifying the analysis by geographic area. Second, we present the results of the application.

A. Data Characteristics

Respiratory cancer is defined to include cancer of trachea, bronchus, lung, and pleura (codes 162 and 163 in Refs. 26-28); however, a recent report indicates that over 98% of these deaths are due to cancers of the bronchus and lung (29). Respiratory cancer mortality has not followed the trend in decline observed for total mortality over the period. For example, while the total age-standardized death rate (ASDR) for white males in the United States declined 15.5% over the period 1968-1977, the ASDR for respiratory cancer increased 20.7% (30, 31).

If the distribution of individual differences in lung cancer risk is fixed at an early age for a cohort, then successive cohorts would exhibit roughly propor-

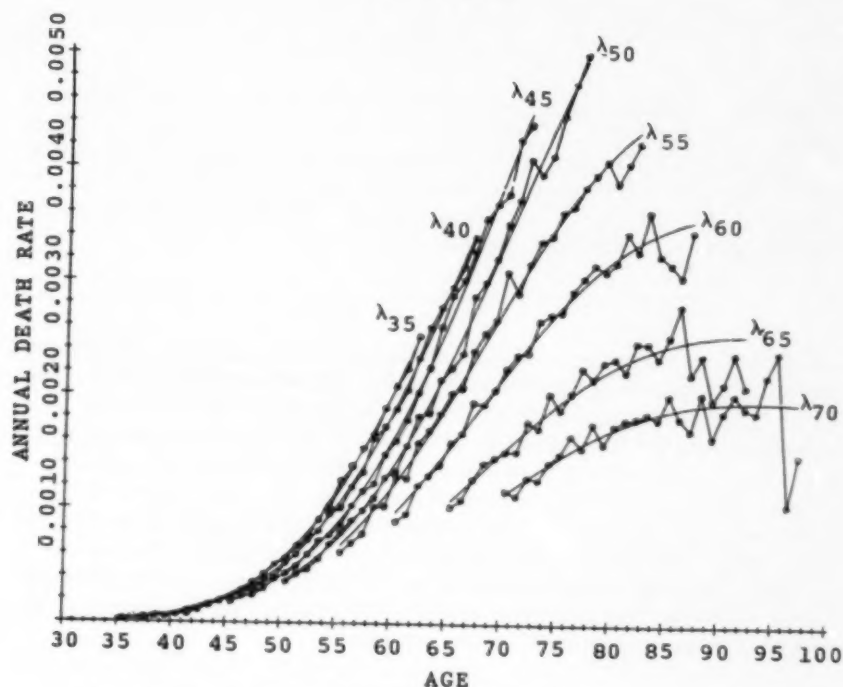


FIG. 1. Lung cancer death rates for eight cohorts of U.S. white males, 1950-1977.

tional increases in the age incidence/mortality curve of lung cancer. To investigate this, we plotted 28 years of death rates observed during the period 1950 to 1977 for each of 8 cohorts which were born 5 years apart starting in 1880 and ending in 1915 (Fig. 1).

The 1880 birth cohort was 70 years of age in 1950 so that the plot for this cohort in Fig. 1 begins at age 70 and ends at age 97. The jagged lines connect the observed death rates while the smooth lines connect the predicted hazard rates under the model with $m = 6$ described below. The 1915 birth cohort was 35 years of age in 1950 so that the plot for this cohort in Fig. 1 begins at age 35 and ends at age 62. Thus, the oldest and youngest cohorts have no ages for which observations are available in both cohorts.

A prior analysis of these cohort lung cancer data (32) examined the statistical fit of a sequence of hierarchical models to determine if an acceptable multistage model could be developed for these data which did not represent distributed susceptibility. The results showed that this was not possible, even when each cohort was modeled with its own mean susceptibility parameter (α_y). (By convention, subscripts on parameters indicate dimensions over which they may vary: y = year of birth, x = age.) Furthermore, though that analysis showed that a common shape parameter (γ) was unacceptable, when the shape parame-

ter was permitted to vary by cohort (γ_y), the estimated values rose from 0.029 (1880 cohort) to 0.116 (1905 cohort) and then dropped to 0.082 (1915 cohort). The finding of increases in γ_y between the 1880 and 1905 cohorts seemed consistent with the expectation that a simple change in proportional hazards between cohorts would be reflected in both α_y and γ_y but not in their ratio, ρ (also, see assumption Eq. [26a]). The finding of a decrease in γ_y from the 1905 to 1915 cohort, however, is inconsistent with the ratio being constant over cohorts. Thus, we investigated the model

$$h(x, y) = \alpha_{x,y}(x - l)^{m-1}/[1 + \rho_y(x - l)^m/m] \quad [34]$$

where all the ρ s in certain subsets of the ρ_y s were constrained to equality. From [34] one can see that the model is parameterized with equality constraints on m and l across cohorts.

Recent evidence from Great Britain has indicated that male lung cancer mortality risks there may have begun to decline (Ref. (33), p. 1298). In order to determine if there was any early indication of such a change for U.S. males the model Eq. [34] was parameterized to allow for period specific changes in risks (i.e., $\alpha_{x,y}$ s), but only in the five youngest cohorts. In addition, cohort-specific models were estimated for 48 of the 50 U.S. states (Alaska and Hawaii excluded) and for the District of Columbia to determine how much between-state heterogeneity contributes to the overall lung cancer pattern. Thus our compartment model was used to assess how cohort, period, and geographic variation in lung cancer mortality risks contribute to recent changes in respiratory cancer mortality rates.

B. Results

We applied our compartment model of the age increases in lung cancer mortality risks within cohorts to determine the extent to which recent changes in respiratory cancer ASDRs can be attributed to cohort differentials in risk. Using integer values of m in the range 5 to 8 we fitted four alternative models to 222 observations for the eight cohorts described above (the last two observations for the 1880 cohort were deleted for consistency of fitting procedures with the analysis of state specific data where age 95 years was the oldest age). In addition to these data we also included shorter follow-up data for five additional cohorts born 5 years apart from 1920 to 1940 with each follow up beginning when the cohort reached age 35 years. These data contributed an additional 65 observations (i.e., $23 + 18 + 13 + 8 + 3 = 65$) to yield a total of 287 observations. The structure of the ρ constraints for each model was as follows: Three values of ρ_y were estimated—one for 1880 and 1885 cohorts, one for the 1890 and 1895 cohorts, and one for the 1900 through 1940 cohorts. Additional structure was imposed on the α s for the five youngest cohorts to allow for period-specific changes in risk.

In Table 1 we see that for all four models there is a consistent pattern in the ρ -estimates with the value for the 1890 and 1895 cohorts lower than for both the earlier and the latter cohorts. The χ^2 values with 259 *df* (i.e., 287 observations

TABLE I
FIT OF FOUR ALTERNATIVE MODELS OF RESPIRATORY CANCER MORTALITY

Model	<i>m</i>	<i>l</i> (years)	<i>p</i>			$\chi^2(259)$
			1880 & 1885	1890 & 1895	1900 & Later	
1	5	25.9	1.13×10^{-8}	0.63×10^{-8}	1.15×10^{-9}	462.9
2	6	20.4	1.98×10^{-10}	1.47×10^{-10}	2.66×10^{-10}	353.5
3	7	14.6	2.57×10^{-12}	2.16×10^{-12}	3.78×10^{-12}	328.8
4	8	8.7	2.68×10^{-14}	2.42×10^{-14}	4.11×10^{-14}	332.0

minus 28 parameters) indicate that the four models are all close to an acceptable statistical fit, a finding which can be verified visually in Fig. 1 for the model with $m = 6$ and $\chi^2 = 353.5$. In Fig. 1 we see that the bulk of χ^2 is due to local lack of smoothness in the observed death rates—not to systematic bias in the fit of the model. The model with $m = 5$ fits substantially poorer than the other three models. This suggests that choice between the models requires additional criteria based on their biological implications.

The fact that all four models require at least three p values suggests that (a) there exists significant period differences in risk factor exposures which differentially affect the risk levels of cohorts who are at susceptible ages during those periods, and (b) the existence of these period effects is independent of the specification of the parameter m . To further interpret these period effects it is necessary to examine parallel changes in risk factor exposures. For example, a variety of exposure factors have been identified for respiratory cancers but one of the most pervasive and strongest is tobacco consumption. We can usefully examine how smoking trends correlate with the period effects if we remain aware of the existence of a range of other exposure factors (e.g., occupation and environmental exposures to toxic substances and asbestos) which also will have contributed to cohort differences and trends.

The 1890 and 1895 cohorts reached age 20 in 1910 and 1915 respectively. Thus, these were the youngest cohorts who fought in World War I. At that time, they were also the first cohorts to be major cigarette consumers differing from the older cohorts who were more likely to be pipe and cigar smokers. Thus, changes in tobacco consumption patterns were probably accompanied by a corresponding change in their age trajectory of lung cancer incidence and mortality. Since a simple proportional change in cohort risk would not require separate p values, we conclude that this change for the 1890 and 1895 cohorts was more complex. For example, with increases in the proportion of smokers, the shape of the susceptibility distribution may have changed significantly.

A second major change occurred beginning with the 1900 cohort which reached age 20 in 1920. In Fig. 1 we see that the later cohort death rates are roughly proportional to those for the 1900 cohort as would be expected since a common p value applies. For all four models, this third p value is the largest

and suggests an increase in the rate of mortality selection for these cohorts. That is, because of the location of ρ in Eqs. [25] and [26b], large ρ values tend to "flatten" out the mortality curve more rapidly than small values.

Though we can plausibly relate the period effects in our model to changes in tobacco consumption it is clear that these changes have a more complex etiology involving a range of occupational and environmental factors. To investigate more fully these dynamics, we maintained the structure of our prior analysis (32) with one α_i value for each cohort, but only for the eight cohorts born in 1880 to 1915 (i.e., the cohorts in Fig. 1). For younger cohorts, with ρ_i constrained equal over the range of birth years from 1900 to 1940, we permitted each cohort to have one $\alpha_{i,y}$ value for each 5 years of cohort follow up and one $\alpha_{i,y}$ value for the final 3 years of cohort follow up in 1975 to 1977. These constraints implied estimation of the 23 α -parameter values presented in Table 2.

In Table 2 all four models show a monotonic increase in α_i from the 1880 to 1915 cohorts, as expected given the increased death rates for younger cohorts observed in Fig. 1. The 1920 to 1935 cohorts show declining α values within each cohort over age groups. In addition, the maximum value of $\alpha_{i,y}$ is obtained for the 1930 cohort in the age interval 35-39 (i.e., in 1965-69) under all four models. These estimates suggest that the lung cancer death rates for younger male cohorts are undergoing a *decline*. This decline is significant in that it parallels a decline in male lung cancer mortality rates in Great Britain. Furthermore, the consistency of the patterns of ρ_i and $\alpha_{i,y}$ under the four models suggests that the observation of the decline is robust to the specification of the parameter m . This suggests that the prior observation (32) of declining γ_i for younger cohorts, which we now are modeling as declines in $\alpha_{i,y}$ from one period to the next, may reflect a third major change resulting from altering cigarette consumption behavior due to warnings (starting in 1962) of the health hazards of tobacco. Such an interpretation follows from Ref. (33), which noted that the declines in tar and nicotine content of today's cigarettes might explain declines in the age-specific lung cancer death rates below age 45 (in 1975).

An important component of any intervention strategy is an accurate assessment of the role of significant risk covariates in generating individual differences in susceptibility to tumor onset. We can illustrate the potential effects of such covariates where covariate functions were assumed to be beta distributed. To do so, we fit the model to state-specific cohort data (for the 48 coterminous states and the District of Columbia) for the nine oldest cohorts (born 1880 to 1920), permitting only the parameter α_i to differ from the national parameter estimates (reported in Ref. (32)). The squared coefficients of variation of these $\hat{\alpha}_i$ s ranged from $\hat{c}_{1910}^2 = 0.0304$ to $\hat{c}_{1880}^2 = 0.0757$, with a median value of $\hat{c}_{1895}^2 = 0.0396$. Thus, we will focus on the 1895 cohort for illustrative calculations.

By matching moments we can determine that the $\hat{\alpha}_{1895}$ s in the 49 areas are proportional to the beta variate $G_1 \sim B(1, 6.21, 2.35)$. As discussed, this is the variate which transforms the variate $Z_0 \sim \Gamma(1, 8.56)$ to the variate $Z_1 = Z_0 \cdot G_1 \sim \Gamma(1, 6.21)$. A plot of this beta density, $q(g_1)$ defined in the range (0, 1.38),

TABLE 2
MAXIMUM LIKELIHOOD ESTIMATES OF α UNDER FOUR ALTERNATIVE MODELS

Age at start of interval	1950-1954	1955-1959	1960-1964	1965-1969	1970-1974	1975-1977
Model 1: $\alpha \times 10^9$						
35	1.62	2.42	3.02	3.66	3.53	3.62
40	1.47	1.62	2.01	2.23	2.52	2.32
45	1.22	1.47	1.62	1.88	2.18	2.11
50	1.04	1.22	1.47	1.62	1.74	1.91
55	0.98	1.04	1.22	1.47	1.62	1.68
60	0.77	0.98	1.04	1.22	1.47	1.62
65	0.54	0.77	0.98	1.04	1.22	1.47
70	0.40	0.54	0.77	0.98	1.04	1.22
75		0.40	0.54	0.77	0.98	1.04
80			0.30	0.54	0.77	0.98
85				0.40	0.54	0.77
90					0.40	0.54
95						0.40
Model 2: $\alpha \times 10^{11}$						
35	2.38	3.01	3.77	4.56	4.40	4.20
40	2.17	2.38	2.88	3.20	3.61	3.26
45	1.75	2.17	2.38	2.79	3.24	3.13
50	1.50	1.75	2.17	2.38	2.58	2.83
55	1.35	1.50	1.75	2.17	2.38	2.47
60	1.05	1.35	1.50	1.75	2.17	2.38
65	0.73	1.05	1.35	1.50	1.75	2.17
70	0.56	0.73	1.05	1.35	1.50	1.75
75		0.56	0.73	1.05	1.35	1.50
80			0.56	0.73	1.05	1.35
85				0.56	0.73	1.05
90					0.56	0.73
95						0.56
Model 3: $\alpha \times 10^{13}$						
35	2.75	3.10	3.88	4.70	4.53	4.16
40	2.51	2.75	3.24	3.60	4.06	3.62
45	2.00	2.51	2.75	3.24	3.77	3.62
50	1.71	2.00	2.51	2.75	3.01	3.31
55	1.51	1.71	2.00	2.51	2.75	2.86
60	1.17	1.51	1.71	2.00	2.51	2.75
65	0.82	1.17	1.51	1.71	2.00	2.51
70	0.62	0.82	1.17	1.51	1.71	2.00
75		0.62	0.82	1.17	1.51	1.71
80			0.62	0.82	0.17	1.51
85				0.62	0.82	1.17
90					0.62	0.82
95						0.62

TABLE 2—Continued

Age at start of interval	1950-1954	1955-1959	1960-1964	1965-1969	1970-1974	1975-1977
Model 4: $\alpha \times 10^{15}$						
35	2.65	2.75	3.43	4.16	4.01	3.60
40	2.41	2.65	3.05	3.39	3.82	3.37
45	1.91	2.41	2.65	3.12	3.63	3.48
50	1.64	1.91	2.41	2.65	2.91	3.20
55	1.42	1.64	1.91	2.41	2.65	2.76
60	1.10	1.42	1.64	1.91	2.41	2.65
65	0.77	1.10	1.42	1.64	1.91	2.41
70	0.58	0.77	1.10	1.42	1.64	1.91
75		0.58	0.77	1.10	1.42	1.64
80			0.58	0.77	1.10	1.42
85				0.58	0.77	1.10
90					0.58	0.77
95						0.58

Note. In all models, parameter estimates for cohorts age 35 and over in 1950 were constrained to be constant over entire interval 1950-1977. The sequence of parameter estimates for each cohort is obtained following diagonally down the table from the entry in the first column or row.

is contained in Fig. 2. One can see that this density is moderately skewed to the left.

Figure 3 contains the theoretical cumulative probability, obtained by integrating the beta density $q(g_1)$, plotted against the observed cumulative probability. Here one can see that the fit of the beta distribution is very good although there is a slight tendency to overestimate the cumulative probability. This bias most likely reflects the use of the method of moments in estimating the parameters, rather than use of an optimal method such as maximum likelihood.

The fitted gamma distribution for this cohort had a shape parameter estimate of $\hat{\gamma} = 0.0691$ which suggests that additional covariate effects are operating beyond the state effects represented in the covariate function G_1 . We can infer the distribution of this additional covariate function if it is also assumed to be of beta form. Letting $Z_2 = Z_1 \cdot G_2$, where $Z_2 \sim \Gamma(1, 0.0691)$ as estimated

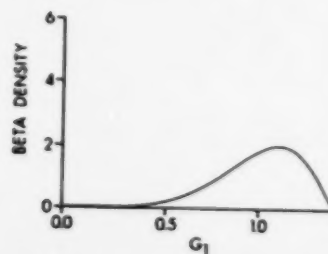


FIG. 2. Beta density, $\delta_1 = 6.2137$, $\theta_1 = 2.3542$.

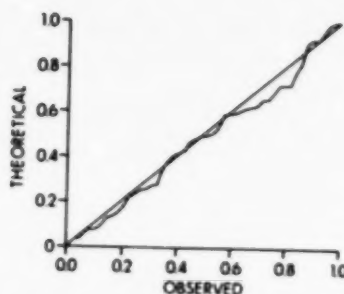


FIG. 3. Theoretical vs observed cumulative probability.

from the national mortality data for this cohort, it follows from the methods described in Section IID that $G_2 \sim B(1, 0.0691, 6.14)$. Plots of the densities of Z_2 and G_2 are contained in Figs. 4 and 5.

One can see that the densities of Z_2 and G_2 are remarkably similar, both functions being strongly skewed to the right, indicating a small proportion of the population with very high relative susceptibilities (e.g., maximum value of G_2 is 89.92). This suggests that the extreme variability of Z_2 is primarily due to the unobserved covariate function G_2 , not to the observed state covariate function G_1 . Indeed, the covariate function G_1 increases the squared coefficient of variation of Z_0 from 0.12 to the Z_1 value 0.16, a 38% increase. In contrast G_2 increases the squared coefficient of variation of Z_1 from 0.16 to the Z_2 value 14.47.

IV. DISCUSSION

Modeling chronic diseases requires the development of statistical methods that recognize that these diseases are processes, evolving over a lengthy period of time, passing through several stages, and operating in a heterogeneous population. We present a statistical model for this analysis based upon a compartment modeling strategy where (a) the probabilistic structure of the process is

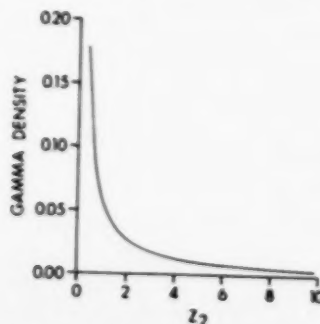


FIG. 4. Plot of gamma density, $\gamma_2 = 0.0691$.

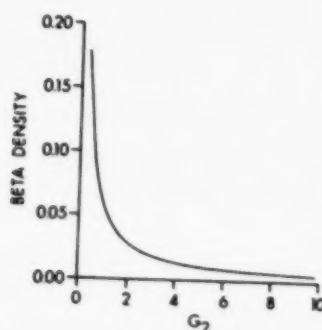


FIG. 5. Plot of beta density, $\delta_2 = 0.0691$, $\theta_2 = 6.1446$.

specified and (b) specific transition and mixing functions are derived from substantive insights to describe the probabilistic component. For this model to be useful it is also necessary to show how risk covariates affect the heterogeneity of the population and to provide a strategy for determining changes in the risk distribution. This is illustrated for the case of a gamma mixing distribution which has a closed form under selection.

The mathematical results are illustrated with cohort data on lung cancer mortality risks from 1950 to 1977. We found evidence that, for recent male cohorts, there is likely to be a decline in the level of mortality risks in the near future (though the actual numbers may increase because of larger cohorts). We also illustrated how the heterogeneity distribution was affected by geographic stratification as a way of representing risk covariate effects.

ACKNOWLEDGMENTS

Dr. Manton's, Mr. Stallard's, and Dr. Woodbury's research was supported by NIA Grant 2 R01 AG01159 and EPA Cooperative Agreement CR807465. Dr. Creason's and Dr. Riggan's research was supported by internal funds of the EPA.

REFERENCES

1. MANTON, K. G., AND STALLARD, E. Maximum likelihood estimation of a stochastic compartment model of cancer latency: Lung cancer mortality among white females in the U.S. *Comput. Biomed. Res.* **12**, 313 (1979).
2. MANTON, K. G., AND STALLARD, E. The use of mortality time series data to produce hypothetical morbidity distributions and project mortality trends. *Demography* **19**, 223 (1982).
3. MANTON, K. G., AND STALLARD, E. A population-based model of respiratory cancer incidence, progression, diagnosis, treatment, and mortality. *Comput. Biomed. Res.* **15**, 342 (1982).
4. COX, D. R. Regression models and life tables. *J. R. Stat. Soc. Ser. B* **34**, 187 (1972).
5. MANTON, K. G., AND STALLARD, E. "Recent Trends in Mortality Analysis." Academic Press, Orlando, Fla., 1984.
6. MANTON, K. G., WOODBURY, M. A., AND STALLARD, E. A variance components approach to categorical data models with heterogeneous cell populations: Analysis of spatial gradients in county lung cancer mortality rates in North Carolina counties. *Biometrics* **37**, 259 (1981).
7. ARMITAGE, P., AND DOLL, R. The age distribution of cancer and a multistage theory of

- carcinogenesis. *Brit. J. Cancer* **8**, 1 (1954).
8. ARMITAGE, P., AND DOLL, R. Stochastic models for carcinogenesis. In "Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV, Biology and Problems of Health" (J. Neyman, Ed.). Univ. of California Press, Berkeley, 1961.
 9. DOLL, R. The age distribution of cancer: Implications for models of carcinogenesis. *J. R. Stat. Soc. Ser. A* **134**, 133 (1971).
 10. WHITTEMORE, A. S. The age distribution of human cancer for carcinogenic exposures of varying intensity. *Amer. J. Epidemiol.* **106**, 418 (1977).
 11. PETO, R., ROE, F. J., LEE, P. N., LEVY, L., AND CLACK, J. Cancer and aging in mice and men. *Brit. J. Cancer* **32**, 411 (1975).
 12. PETO, R. Epidemiology, multistage models, and short-term mutagenicity tests. In "Origins of Human Cancer" (H. H. Hiatt, J. D. Watson, J. A. Winston, Eds.), pp. 1403-1428. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1977.
 13. COOK, P. J., DOLL, R., AND FELLINGHAM, S. A. A mathematical model for the age distribution of cancer in man. *Int. J. Cancer* **4**, 93 (1969).
 14. DOLL, R., AND PETO, R. Mortality in relation to smoking: 20 Years' observations on male british doctors. *Brit. Med. J.* **2**, 1525 (1976).
 15. DOLL, R., AND PETO, R. Cigarette smoking and bronchial carcinoma: Dose and time relationships among regular smokers and lifelong non-smokers. *J. Epidemiol. Community Health* **32**, 303 (1978).
 16. DAY, N. E., AND BROWN, C. C. Multistage models and primary prevention of cancer. *J. Natl. Cancer Inst.* **64**, 977 (1980).
 17. BURCH, P. R. J. "The Biology of Cancer: A New Approach." Univ. Park Press, Baltimore, 1976.
 18. Burch, P. R. J. Smoking and lung cancer: The problem of inferring cause. *J. R. Stat. Soc. Ser. A* **141**, 437 (1978).
 19. BEARD, R. E. A theory of mortality based on actuarial, biological and medical considerations. In "International Population Conference, New York, 1961, (Proceedings)", Vol. 1, p. 611. International Union for the Scientific Study of Population, London, 1963.
 20. CHIANG, C. L. "An Index of Health: Mathematical Models. Vital and Health Statistics, Data Evaluation and Methods Research." DHEW (PHS) Publ. 1000—Series 2—No. 5, Washington, D.C., 1965.
 21. SHEPARD, D., AND ZECKHAUSER, R. "Interventions in Mixed Populations: Concepts and Applications." J. F. Kennedy School of Government Discussion Paper Series 49D. Harvard University, 1977.
 22. MCGILCHRIST, C. A., AND SIMPSON, J. M. Effect of differential mortality rates on risk-variable distribution. *Math. Biosci.* **43**, 173 (1979).
 23. VAUPEL, J. W., MANTON, K. G., AND STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439 (1979).
 24. HOUGAARD, P. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **71**, 75 (1984).
 25. MANTON, K. G., STALLARD, E., AND VAUPEL, J. W. Alternative models for the heterogeneity of mortality risks among the aged. *J. Am. Stat. Assoc.*, to appear.
 26. World Health Organization. "International Classification of Diseases: Manual of the International Statistical Classification of Diseases," 6th revision. World Health Organization, Geneva, 1948.
 27. World Health Organization. "International Classification of Diseases: Manual of the International Statistical Classification of Diseases," 7th Revision. World Health Organization, Geneva, 1957.
 28. World Health Organization. "International Classification of Diseases: Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death," 8th Revision. World Health Organization, Geneva, 1969.
 29. National Technical Information Service. "Multiple Causes of Death Public Use Tape: ICDA-8 (1968-1978)." U.S. Dept. Commerce, National Center for Health Statistics, 1980.

30. KLEBBA, A. J. "Mortality from Diseases Associated with Smoking: United States, 1950-64." National Center for Health Statistics, Series 20, No. 4, PHS Publ. 1000, Washington, D.C., 1966.
31. KLEBBA, A. J. "Mortality from Diseases Associated with Smoking." Vital and Health Statistics: Series 20, Data from the National Vital Statistics System, No. 17, DHHS Publ. (PHS) 82-1854, 1982.
32. MANTON, K. G., STALLARD, E., AND RIGGAN, W. B. Strategies for analyzing ecological health data: Models of the biological risk of individuals. *Stat. Med.* 1, 163 (1982).
33. DOLL, R., AND PETO, R. The causes of cancer: Quantitative estimates of avoidable risk of cancer in the United States today. *J. Natl. Cancer Inst.* 66, 1191 (1981).

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.